DOCUMENT RESUME

ED 417 215 TM 028 158

AUTHOR Woldbeck, Tanya
TITLE Basic Concepts in Modern Methods of Test Equating.
PUB DATE 1998-04-11
NOTE 24p.; Paper presented at the Annual Meeting of the Southwest Psychological Association (New Orleans, LA, April 1998).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Data Collection; Difficulty Level; *Equated Scores; *Item Response Theory; Test Construction; *Test Format
IDENTIFIERS Equipercentile Equating; Item Discrimination (Tests); *Linear Equating Method; Vertical Equating

ABSTRACT

This paper summarizes some of the basic concepts in test equating. Various types of equating methods, as well as data collection designs, are outlined, with attempts to provide insight into preferred methods and techniques. Test equating describes a group of methods that enable test constructors and users to compare scores from two different forms of a test. Horizontal equating is performed between two different versions of a test, and vertical equating is performed on tests across difficulty levels. The most basic of the equating methods is linear equating, which assumes that the two tests to be equated differ only in means and standard deviations. Equipercentile equating considers scores to be equivalent if the percentile ranks corresponding to the scores on two forms of a test are equal. Item response theory equating is a viable alternative to more conventional methods of equating. It uses item characteristic curves to describe the relationship between a score on a test and the item difficulty and item discrimination. While it is beyond the scope of the paper to go into great detail, some issues related to test equating are considered. More detailed readings are recommended. (Contains 2 figures and 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 417 215

Running Head: MODERN METHODS IN TEST EQUATING

Basic Concepts in Modern Methods of Test Equating

Tanya Woldbeck

Texas A&M University 77843-4225

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Tanya Woldbeck

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028158

Paper presented at the annual meeting of the Southwest Psychological Association, New Orleans, April 11, 1998.

BEST COPY AVAILABLE

Abstract

The present paper summarizes some of the basic concepts in test equating. Various types of equating methods as well as data collection designs will be outlined, with attempts to provide insight into preferred methods and techniques. While it is beyond the scope of the current paper to go into great detail, some issue related to test equating will be considered. The reader will be introduced to the area of test equating from a very general perspective. More detailed readings will also be recommended.

Basic Concepts in Modern Methods of Test Equating

In the area of test construction, it has been emphasized that careful inspection of the content and objectives forms the beginning of good testing practice. Large testing programs, especially, are held to a high degree of scrutiny in their testing practices. In the context of large-scale, high-stakes tests, such as the SAT, testing programs must consider various legal, psychometric and practical issues on a daily basis.

One such issue is the necessity of creating alternate forms of a test. It is unsound practice to administer the exact same test on any two separate occasions. This practice results in breaches of security, giving the latter group a distinct advantage over the former group. But it is legally offensive to administer different forms of the same test to different people unless the tests can be shown to be equivalent. In addition, being human, test constructors cannot create two forms of a test, with different questions, and expect them to be parallel in all other respects (i.e., provide the same score for any given examinee), and therefore formulas must be applied to make scores from alternate forms equivalent.

In an ideal psychometric world, all constructed test forms would be strictly parallel, all samples of test-takers would be completely randomly selected and equal in ability and laws and ethics would not restrict testing programs in any way. These conditions are seldom met, however. All of these issues facing formal testing programs, as well as many more, make it necessary to utilize what is known as test equating procedures.

This paper will first define test equating, and introduce the major assumptions underlying the various methods. Data collection designs will then be explained to lay the foundation for the practical use of test equating. This discussion will then turn to describing

the basic steps and procedures for three major types of test score equating: (a) linear procedures, (b) equipercentile procedures, and (c) those procedures born from the concepts of item response theory (IRT). Following will be a brief discussion of which methods to use, and some issues associated with the various test equating techniques.

Test equating describes a group of methods which enable test constructors and users to compare scores from two different forms of a test. Definitions of test equating vary widely from one author to the next. Lord (1980) offered the following, conceptual definition:

> If an equating of tests x and y is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level θ whether they are to take test x or test y. (p. 195)

Angoff (1971), in his extensive treatment of conventional equating methods, described equating as the conversion of "the system of units of one form to the system of units of the other - so that scores derived from the two forms *after* conversion will be directly equivalent" (p. 562). In their introduction to the topic, Crocker and Algina (1986) differentiated between horizontal and vertical equating. *Horizontal equating* is performed between two different forms of a test. An example would be equating various forms of the GRE as they are administered across a five year period; it s still important that the scores are reasonably comparable across forms and time. Vertical equating relates to scores on two different levels of a test. *Vertical equating* would be performed on various standardized achievement tests across difficulty levels (i.e grade levels) (Crocker & Algina, 1986). For example, it may be helpful to be able to interpret standardized test scores for a child taken in

each year of elementary school. While the complexities of vertical equating will be discussed later in the paper, most of the current discussion will be focused on horizontal equating.

The basic purpose of test equating, simply, is "to establish, as nearly as possible, an effective equivalence between test scores" (Petersen, Marco, & Stewart, 1982, p. 72). The process, which is not always so simple, is to determine an equating function f(x) to "map the raw scores obtained from a newer test form into raw scores obtained from an older test form" (Braun & Holland, 1982, p. 12). In essence, equating procedures are numerical adjustments made to scores on two forms of a test (Braun & Holland, 1982).

Angoff (1971) likened this numerical conversion to the conversion formula used to compare degrees Fahrenheit to degrees Celsius. The formula to convert measured temperature from the Celsius scale ° to the Fahrenheit scale (F) is as follows:

$$C = (F - 32) / 2$$

Likewise, the conversion can algebraically be inverted in the opposite direction to determine degrees Fahrenheit from Celsius:

$$F = (C * 2) + 32$$

This example will be used again to illustrate points made later in the paper.

If equating has been successful, it is possible to discuss true growth over several administrations of a test, true changes in a populations performance over a given period of time, and to compare students who take tests at various times during the year (Angoff, 1971). As with most statistical procedures, there are assumptions inherent in the process of test equating. The most basic of these assumptions is that the two tests to be

equated must measure the same characteristic. A test of reading can, theoretically, only be equated with another test measuring reading. To use Angoff's (1971) example, it would seem silly to equate degrees Fahrenheit with weight in pounds.

Along these same lines, it is important that the two tests to be equated be *unidimensional*. Not only should the tests measure the same thing, but they should measure only one thing. Holmes (1986) suggested that with a multidimensional test like an achievement test covering more than one content area, equating should be done on a subtest level. The notion of unidimensionality becomes especially important in the discussion of equating with item response theory (IRT) methods.

A concept which is necessary to prove unidimensionality is local independence. Local independence requires that at any given ability level, θ, responses to two items on a test are independent of one another. In other words, an examinee's response on one question is completely unrelated to his or her response on another item. Across ability levels, in contrast, we want the items to be dependent of one another, unified by a single latent construct, which is the concept known as unidimensionality (Crocker & Algina, 1986).

Another assumption states that the conversion function must be independent of the sample from which it was calculated (Angoff, 1971). This is closely related to the concept of invariance in IRT. In addition, the conversion function must be invertible: "A basic requirement of equating is that the result should be the same no matter which test is called x and which is called y" (Lord, 1980, p. 198). It may seem obvious to the reader at this point that, due to many of the mentioned assumptions, regression, or prediction, is not the same as equating. As Lord (1980) noted, these requirements are "not satisfied when we predict

one test from another" (p. 198), and "if regression methods of "equating" are used....examinees could properly complain." (p. 207). It would, in this case, make a difference which test an examinee took. While some of the equations and terms used in equating may seem similar to those used in regression, the similarities end there.

While not considered formal assumptions, other authors (Crocker & Algina, 1986; Holmes, 1986) have argued that the two tests to be equated should be of similar difficulty and yield equally reliable scores. While Angoff (1971) has provided the interested reader with methods of equating two tests which do not produce equally reliable scores, he cautions against this practice, stating that the "scores cannot be 'equated' in any meaningful way" (p. 571). If this were allowed, there would be no need to create tests which yield reliable test scores. We would need only to equate the test with unreliable sores to one which does yield reliable scores. These assumptions will be discussed again later, when deciding upon an appropriate design for each individual research purpose.

Four Designs

In order to equate two tests, the test constructor must first decide between a number of data collection designs. The number of designs described in the literature range from three (Crocker & Algina, 1986) to five or more (Angoff, 1971). This paper will describe four of these designs, appropriately named Designs I, II, III, and IV (Angoff, 1971). The first three designs all assume random assignment to groups. These designs are described in greater detail in Angoff (1971), Crocker and Algina (1986) and in Holmes (1986).

Design I, also called an equivalent group design (Holmes, 1986), entails administering two forms of a test (X and Y) to two groups of examinees. The two groups are formed from

BEST COPY AVAILABLE

a larger population by random assignment. A single-group design (Design II), administers both forms to both groups of examinees in a counterbalanced order to control for any practice effect.

The third design administers one form to each group, just as in Design I. However, each group also takes a set of common items, or an anchor test, which is used to equate the two forms. This anchor test must measure what the two tests measure, and can be internal or external. An internal anchor test is administered along with the tests to be equated and contributes to the overall score for each examinee. An external anchor test is separately timed and administered, and is not part of an examinee's total score (Crocker & Algina, 1986).

The anchor-test design seems to be the most popular design in the research literature, and there have been studies conducted to determine the appropriate length of the anchor test. Angoff (1971) suggested as a rule of thumb that the anchor test contain at least 20 items or be at least 20% as long as the tests to be equated, whichever number is larger. Budescu (1985) cautioned against blind reliance on this rule of thumb, and suggested that the length of the anchor test should reflect the reliability of the scores yielded from the tests to be equated. The anchor test should be longer for tests which yield less reliable scores. Cook and Petersen (1987) reviewed some of the research and concluded that "the properties of the anchor test can seriously affect conventional equating results" (p. 238). These authors further stated that the properties of the anchor test become less important "as the equating tests become more similar in level and dispersion of ability" (p. 238). This is because an important role of the anchor test is to control for differences in the two groups of examinees

taking two different forms of the test. As Lord (1980) noted

> [D]ifferences between... two samples of examinees can be measured and controlled by administering to each examinee an anchor test measuring the same ability as x and y. When an anchor test is used, equating may be carried out even when the x group and the y group are not on the same ability level (p. 200)

This special aspect of the anchor test leads to Design IV which specifically considers the more common situation where random assignment to test form is not possible or practical. Consider the Graduate Record Exam (GRE) as an example. It is possible to take the GRE at various times throughout the year. Some are ready to take the test earlier than others, some procrastinate on deadlines, and end up taking the test at later dates. Inherently, these groups of people are going to differ beyond random sampling error. Design IV techniques look to minimize these differences. Angoff (1971) gives an excellent description of these techniques.

Selecting A Design

The challenge becomes which design to choose. This decision is based on a number of variables including how long the test is, how long the examinees have to take the exam, how different the groups are, and what the legal and ethical ramifications are. This discussion will focus on the drawbacks of each design.

In Design I, the two forms are given to two different groups. There are, therefore, no items which have been answered by more than one person. This aspect increases the amount of equating error. In fact, to contain as little error as Design II, one would need 10 times as many examinees in Design I (Crocker & Algina, 1986). Therefore, Design I will

BEST COPY AVAILABLE

be accurate only if a large number of examinees are available. There would, of course, be legal and security issues to be considered. Half of the people taking the test would be taking an old test. Given a high stakes test like the GRE, that would probably matter to most test-takers.

In Design II, while the error of equating would decrease due to everyone taking both tests, practical issues would complicate matters. If the two tests to be taken are long, it would mean a very long testing time for examinees. Fatigue would become an issue. Again, there would be legal and issues to consider (Homes, 1986).

Designs III and IV seem to alleviate some of the issues raised thus far. The anchor test would not add as much to testing time as Design II, and there would be common items to control for differences between the two groups of examinees. The anchor-test design is the most promising design to date, and is found most often in the literature (Crocker & Algina, 1986; Holmes, 1986). Angoff (1971) adds to this list of designs, and the interested reader is encouraged to consult his discussion.

Having given a background on the methods of collecting data to perform test equating, it is now appropriate to move on to a discussion of types of equating procedures. While linear equating, equipercentile equating, and IRT equating all have the same ends, their means are quite different.

Linear Equating

The most basic of the equating methods is linear equating. Linear equating assumes that the two tests to be equated differ only in means and standard deviations. The distributions of raw scores for the two tests have to be equal, however. Crocker and Algina

(1986) define equivalent scores as those that "can be identified by determining the pair of scores, one on form X and one on form Y, that have the same z-score" (p. 458). The conversion from one test to another is accomplished using additive and multiplicative constants in the form of the following equation (Angoff, 1971):

$$Y = AX + B$$

This equation is used for all of the designs outlined above, the only difference being the calculations of A and B, which can be found in Angoff (1971) and in Crocker and Algina (1986). This linear function works because of the assumption that, beyond means and standard deviations, all other moments are equal. Additive and multiplicative constants only affect the mean and standard deviation of a distribution of scores.

As mentioned, linear equating is only appropriate when the raw score distributions are equal except for means and standard deviations. In Design IV, however, random assignment is not necessary, and standard linear equating would not be appropriate. Angoff (1971) outlines several new techniques which overcome this problem, and facilitate wider use of linear equating in the literature. Cope (1987), MacCann (1989, 1990), and Woodruff (1989) offer more detailed descriptions of these techniques and which are more appropriate in a given research design. Tucker's equally reliable method, Levine's equally reliable method, alternative methods suggested by Angoff (1971), and combinations thereof, are but a few of the methods described in these papers. All of the papers discuss equating situations where the two groups differ in ability and an anchor-test design was used to gather equating data.

Woodruff (1989) looked at modifications of the Tucker and Levine methods. Woodruff also considered the situation where the anchor test is not as closely correlated with

tests as would be desirable based on an earlier stated assumption. He concluded that an Angoff-modified Levine method "was more sensitive than Tucker's method" (p. 260) in this case. Woodruff also described a modification of the Levine method (the Congeneric-Levine method) as "an appealing alternative...because it permits large group difference (as determined by the performance of the two groups on the anchor test)" (p. 260). This conclusion offers great hope for use of these methods in practical testing situations.

Cope (1987) compared the Tucker and Levine Equally Reliable methods with alternative (Design V) methods outlined by Angoff (1971). Cope found similar findings across five different equating methods altogether, and concluded that the Angoff alternative methods showed promise in practical situations due to fewer restrictive assumptions than the Tucker and Levine methods. Technical aspects are far beyond the scope of the current paper, and the interested reader is referred to Angoff (1971) and the papers mentioned in the previous paragraphs for a more detailed discussion.

In summary, if one believes that the raw score distributions of their two test differ only in mean and standard deviation, and that both tests yield equally reliable scores, then linear equating is quite appropriate. Linear equating is quite simple and very practical to use. If all of these restrictions are not met, one may employ the methods outlined by Tucker and Levine, or consider the viability of equipercentile equating.

Equipercentile Equating

Equipercentile equating considers scores to be equivalent if the percentile ranks corresponding to the scores on two forms of a test are equal. Percentile ranks for two sets of scores are compared in order to make the two cumulative d stributions look the same

BEST COPY AVAILABLE

(Cook & Petersen, 1987). Crocker and Algina (1986) outline the steps for equipercentile equating.

The first step is to calculate percentile ranks for each score in each distribution of interest. These percentile values can be plotted against the raw score values to obtain two ogives like the ones seen in Figure 1. There is one curve for each instrument to be plotted. The curves are constructed in a number of ways, by hand or analytically. Smoothing by hand simply involves

connecting the data points. Analytic smoothing is more complicated and is outlined in Angoff (1971) and Cook and Petersen (1987). Fairbank (1987) describes pre-smoothing versus post-smoothing techniques. Pre-smoothing is performed on the cumulative distributions, seen in Figure 1, while post-smoothing techniques are performed after equating, on the single ogive seen in Figure 2.

Insert Figures 1 and 2 about here

Figure 2 is obtained after equivalent scores are determined from the two ogives in Figure 1. Using the two ogives found in Figure 1, corresponding score values are found which have the same percentile rank. These scores are then plotted to construct the curve found in Figure 2 (Crocker & Algina, 1986). Analytic smoothing techniques are beyond the scope of the current paper, however, it should be noted "that smoothing may significantly increase the robustness of equating results, particularly when sample sizes are small" (Cook & Petersen, 1987, p. 227-228).

Problems occur with smoothing, and equipercentile equating in general, which contrast some of its advantages. Equipercentile equating is helpful when the distributions are only slightly different, as in vertical equating. These procedures make fewer assumptions than linear equating, and can therefore be used in a wider variety of research and testing situations.

However, if the distributions are extremely disparate, equating becomes meaningless. In addition, equipercentile equating is a very data-dependent method which simply "compresses and stretches the score units on one test so that its raw score distribution coincides with that on the other test" (Cook & Petersen, 1987, p. 226). This is very problematic when the sample sizes are so small that extreme scores are under represented, and curves are irregular and step-like. This makes it very difficult to replicate the equating in future samples of the same size (Cook & Petersen, 1987). It is difficult to overlook the increased error involved in smoothing, and the fact that equipercentile equating is more complicated than linear equating. An equipercentile equating" of raw scores [does have]... the [additional] convenient property that when a cutting score is used, the proportion of selected examinees will be the same for those taking test X and for those taking test Y, except for sampling fluctuations (Lord, 1980, p. 207).

And as will be discussed later, equipercentile equating even produces more stable findings than some Item Response Theory (IRT) methods (Skaggs & Lissitz, 1986), which will be discussed in the next section.

While the equipercentile method in an anchor test design is very complicated, and better outlined in Angoff (1971), in Design II, when two groups take both forms of the test,

BEST COPY AVAILABLE

equipercentile equating is similar to the procedure outlined above. The only difference between the designs is that data for each test form are aggregated across groups (Angoff, 1971).

Methods Using Item Response Theory

In the more common case of Designs III and IV, where the assumptions of linear and equipercentile equating may not be met, IRT equating has emerged (as it has in many other areas) as a viable alternative to the more conventional methods. IRT uses item characteristic curves to describe the relationship between a score on the test and the item difficulty, item discrimination, and may also include a guessing component (three-parameter model). The basic assumption is that one single latent trait underlies a persons score on a test (Crocker & Algina, 1986). IRT attempts "to model an examinee's performance on a test item as a function of characteristics of the item and the examinee's ability on some unobserved, or latent, trait" (Skaggs & Lissitz, 1986b, p. 497). In any given IRT model (e.g., Rasch model or three-parameter model), the item characteristics and latent trait are expressed in logits. The average latent trait is arbitrarily set at zero logits. Once this is accomplished, the only difference between two forms of a test is the scale. The average ability arbitrarily set at zero is different for each form. However, now, the equating process is simply a linear transformation using the following equation:

$$\theta^* = \theta + m,$$

where θ^* is the transformed ability level and m is a constant use to place the two values on the same scale. The only difficulty is that we never really know someone's true ability level θ. Recall, however, that item characteristics are also scaled the same way (i.e., logits) as

BEST COPY AVAILABLE

ability levels on any given test form. We can calculate item characteristics like item difficulty (b), for example. The equation know becomes:

$$b^* = b + m.$$

We can then easily solve for m and add this value for each estimated ability level on a test to be equated (Crocker & Algina, 1986; Skaggs & Lissitz, 1986b). These new values become θ^*, the equated ability levels.

In general, IRT models behave quite inconsistently (Petersen, Cook, & Stocking, 1983). Forsyth (1981) reminds us that the Rasch, or one-parameter, model, is very restrictive in its assumptions, only taking into consideration the difficulty of an item. On tests which allow for guessing, the Rasch model is completely inappropriate (Skaggs & Lissitz, 1986b). However, when the data fit the Rasch model, it is a very tight and stable equating procedure (Skaggs & Lissitz, 1986). The three-parameter model has shown to be promising because it seems more comprehensive, however, research shows the process to be very effected by sample size (Skaggs & Lissitz, 1986b).

Vertical Equating

Vertical equating presents other issues which make equating levels of an achievement test, per say, quite difficult. Across levels of an Achievement test, the content becomes more and more difficult as the levels increase. This change in difficulty may be so great that at very high levels, what is actually tested is quite different than what is tested at a lower level. The two tests may actually look very different. Lloyd and Hoover (1980) warn that extreme caution should be used in vertical equating because the content specifications

BEST COPY AVAILABLE

normally change dramatically across different levels of an achievement test. In this respect, linear equating is completely inappropriate. The two distributions cannot be the same.

Skaggs and Lissitz (1986) further add that the Rasch IRT model is inadequate in vertical equating due to an inability of the one-parameter model to capture the influence of guessing and item discrimination, which also change across levels of an achievement test. Slinde and Linn (1978, 1979) have also cautioned against reliance on the Rasch model in extreme situations (i.e., large differences in difficulty between test levels). Surprisingly, Skaggs and Lissitz (1986) also stated, based on a review of the literature that the three-parameter model, while better than the one-parameter and linear models, did not consistently perform better than equipercentile equating.

Making the Right Choice

Given that the literature has many different conclusions to draw regarding the various designs and equating methods, which choice needs to be made becomes very situation dependent. Large-scale testing programs will make very different decisions when equating. Given the very large sample sizes programs like ETS have available to them, error becomes less of an issue, but security has a large role. In a small classroom setting, accuracy in equating may not be great, however, simplicity is a necessity. When choosing the right combination for you, setting, sample size, security, tolerance for error, and resources must all be taken into consideration when choosing the proper procedure. IRT methods require computer programs which estimate parameters, and these programs can be expensive.

This paper has outlined several alternatives test users have when alternate forms are necessary to construct. Other sources of information have been provided, and other issues

BEST COPY AVAILABLE

have been summarized. However, all of the authors reviewed for this paper, agree that while equating has been around for quite some time, there is still much to be learned regarding the science of test score equating.

References

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.) Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Braun, H.I. & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D.B. Rubin (Eds.), Test equating (pp. 9-50). New York: Academic Press.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Cook, L.L. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Cope, R.T. (1987). How well do the Angoff design V linear equating methods compare with the Tucker and Levine methods? Applied Psychological Measurement, 11, 143-149.

Holmes, S.E. (1986). Test equating and credentialing examinations. Evaluation and the Health Professions, 9, 230-249.

Lloyd, B.H. & Hoover, H.D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Lord, F.M. (1980). Applications of Item Response Theory to practical testing problems. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

MacCann, R.G. (1989). A comparison of two observed-score equating methods that

assume equally reliable, congeneric tests. Applied Psychological Measurement, 13, 263-276.

MacCann, R.G. (1990). Derivations of observed score equating methods that cater to populations differing in ability. Journal of Educational Statistics, 15, 146-170.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Petersen, N.S., Marco, G.L., & Stewart, E.E. (1982). A test of the adequacy of linear scores equating models. In P.W. Holland & D.B. Rubin (Eds.), Test equating (pp. 71-136). New York: Academic Press.

Skaggs, G. & Lissitz, R.W. (1986). An exploration of the robustness of four test equating models. Applied Psychological Measurement, 10, 303-317.

Skaggs, G. & Lissitz, R.W. (1986b). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Slinde, J.A. & Linn, R.L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-35.

Slinde, J.A. & Linn, R.L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 16, 159-165.

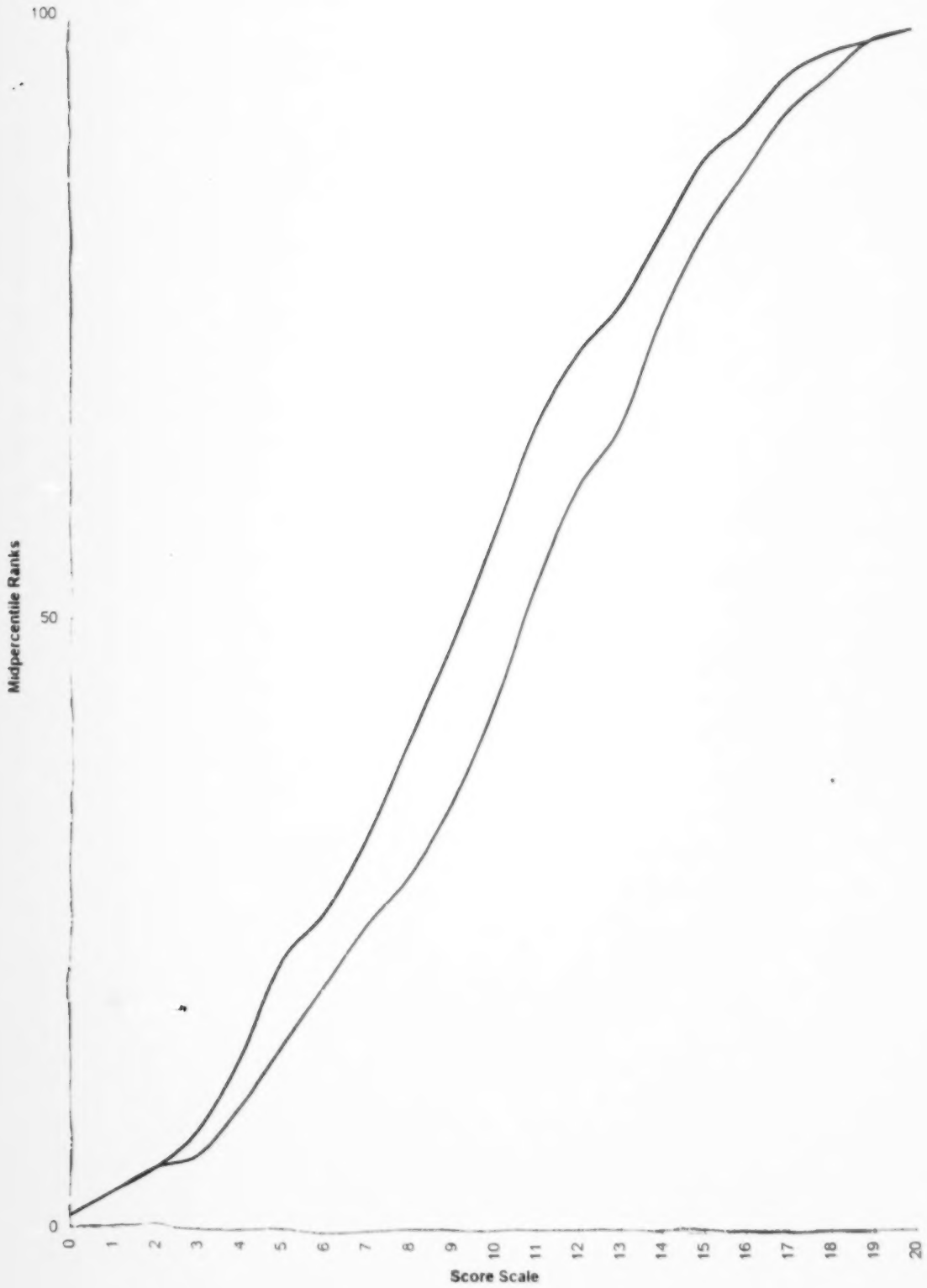
Woodruff, D.J. (1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design. Applied Psychological Measurement, 13, 257-261.

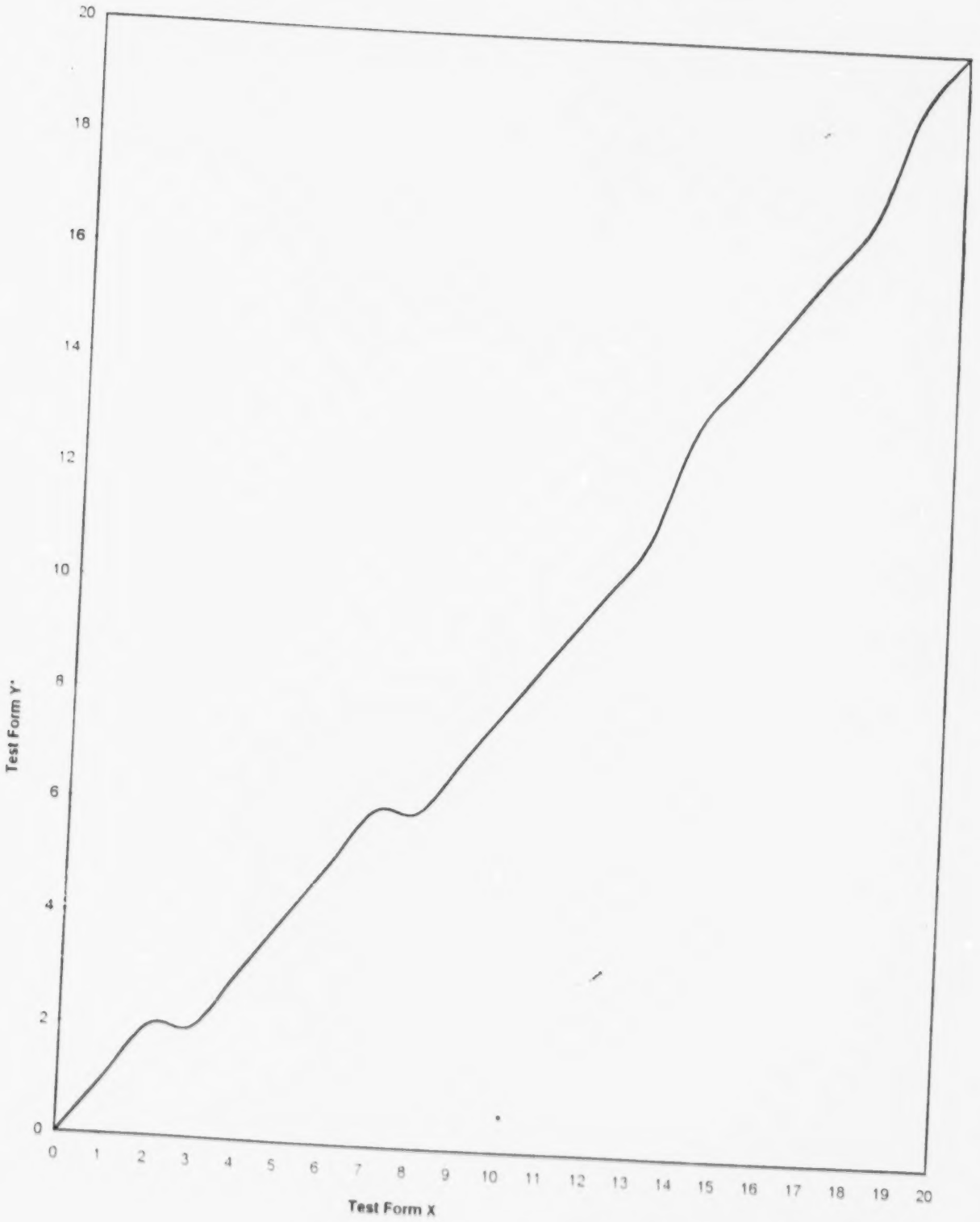
BEST COPY AVAILABLE

Figure Captions

Figure 1. Plot of percentile ranks for two hypothetical 20-item instruments.

Figure 2. Plot of equipercentile equivalent scores for two hypothetical 20-item instruments.


100
50
0
Midpercentile Ranks
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Score Scale



BEST COPY AVAILABLE